Wisdom of the (expert) crowd: Performance of Aggregation Models for Fetal Heart Rate Judgments

Medhini Urs
Department of Psychology, Stony Brook University, NY 11794, USA

Christian C. Luhmann
Department of Psychology, Stony Brook University, NY 11794, USA

Abstract

Existing work has attributed the low clinical utility of continuous CTG to the reliability and accuracy of Cardiotocography judgments. The aim was to determine whether aggregating the judgments of multiple obstetricians (leveraging the "wisdom of crowds") using obstetricians' optimized estimates of the probability of hypoxia improves accuracy. In the current study, we apply three different aggregation techniques to the evaluations of nine obstetricians from the CTU-CHB Intrapartum Cardiotocography Database. The evaluations were optimized estimates of the probability of hypoxia in each evaluation category. The three aggregation models ranged in complexity from an unweighted aggregation scheme to an approach that weighted evaluations based on the contribution of the obstetricians. All the aggregation models were shown to improve judgment accuracy above chance performance. However, the most accurate model was the one which calculated the simple average of obstetricians' judgments. There was no additional benefit of selecting obstetricians who were positive contributors to an ensemble and weighing the evaluations based on their contribution to the ensemble. Aggregating obstetricians' evaluations may be a solution to the ongoing reliability and accuracy issues in fetal heart rate judgment.

Keywords: cardiotocogram (CTG) judgment; obstetrician decision-making; aggregation models; wisdom of crowds

Introduction

Cardiotocography (CTG) is one of the most commonly used medical tools for monitoring vital fetal physiological information, with approximately 89% of women who go into labor monitored via cardiotocography (Declercq et al., 2014). CTG provides intricate information about the fetal heart rate (FHR) and uterine activity during labor (O'Brien-Abel, 2020; Windrix, 2017) and is used to identify babies at risk of hypoxia, an oxygen deficiency that can lead to brain damage and lifelong disabilities, including cerebral palsy. If the risk is great enough, babies may need to be delivered via operative means (e.g., cesarean delivery). Despite widespread use, there is limited evidence for the clinical value of continuous CTG, (Graham et al., 2006; Grant et al., 1989; Nelson et al., 1996; Shy et al., 1990) though it may be beneficial in certain high-risk pregnancies (Pattison & McCowan, 2004). Conversely, the use of continuous CTG substantially increases rates of caesarean sections which is associated with a higher risk of infection and complications in future pregnancies (Alfirevic et al., 2017). Thus, the use of continuous CTG may be, at best, of little clinical use and, at worst, put mothers and babies at increased risk.

The primary explanation for the poor clinical value of CTG is that clinicians' CTG evaluations are inconsistent (Palomäki et al., 2006). Past research has demonstrated that both inter-observer agreement and intra-observer agreement are poor (Bernardes et al., 1997; Blackwell et al., 2011; Hruban et al., 2015; Spilka et al., 2014). This lack of reliability has been shown in both obstetricians (Donker et al., 1993; Nielsen et al., 1987) and midwives (Blix et al., 2003). Furthermore, national and international standards for CTG evaluation (American College of Obstetricians, Gynecologists, et al., 2010; Ayres-de-Campos et al., 2015) provide guidance about features of CTG recordings (e.g., baseline variability of the fetal heart rate) and how the relevant CTG features themselves are defined (e.g., how is baseline heart rate determined?). These guidelines are also associated with diagnostic categories (e.g., normal, suspicious, pathological). Nonetheless, clinicians still fail to agree on both the overall classification of fetuses (Ayres-de-Campos et al., 1999) and the presence/absence of individual features of the CTG recordings (Devane & Lalor, 2005; Sholapurkar, 2015; Todros et al., 1996).

One strategy for dealing with unreliable judgment is to aggregate the judgments of multiple individuals into a single, ensemble judgment, leveraging the so-called ``wisdom of crowds" (Galton, 1907; Hertwig, 2012; Surowiecki, 2005). Aggregating will tend to reduce the noise present in individual judgments and emphasize the information common across the individual judgments (Satopää et al., 2014). Intuitive aggregation schemes such as majority voting and simple averaging are straightforward, but there are more sophisticated schemes (Dhami et al., 2015; Molteni et al., 1996; Tetlock & Gardner, 2016; Toth & Kalnay, 1993).

In medical settings, researchers have investigated a variety of ways to aggregate medical decisions (Barnett et al., 2019; Blutinger et al., 2021; Green et al., 2008; Hermann et al., 2019; Poses et al., 1990; Kämmer et al., 2017; Kattan et al., 2016); Krockow et al., 2020; Kurvers et al., 2016). In the case of continuous evaluations (e.g., treatment duration, survival probability), techniques have included taking the mean evaluation (Kattan et al., 2016; Poses et al., 1990) and then taking the median evaluation (Krockow et al., 2020). In the case of categorical evaluations (e.g., malignant vs. benign), techniques have included taking the most common judgment within the ensemble (i.e., majority vote (Bernstein et al.,

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

2011) and taking the judgment of the most confident clinician (i.e., the confidence rule (Kurvers et al., 2016). Indeed, one approach has been a majority voting scheme to aggregate experts' judgments of continuous CTG (Spilka et al., 2014) (though it has not been established whether doing so improves clinical evaluations).

One drawback of these simple aggregation schemes is that they consider the judgment of each individual to be equally useful. As a result, a larger number of low-ability individuals within the crowd can ultimately ``overrule'' a smaller number of high-ability individuals. There is evidence that this exact scenario limited the performance of ensembles making cancer diagnoses (between three and five radiologists and dermatologists) (Kurvers et al., 2016). Ensembles constructed via majority vote outperformed the best clinician only when the ability of the constituent clinicians was relatively similar. Otherwise, the lower-ability clinicians lower the performance of the ensemble below the performance expected from the higher-ability clinicians.

To deal with this difficulty, researchers have developed techniques to construct weighted ensembles, such that each individual contributes to the ensemble in proportion to their assigned weight. For example, the contribution weighted model (CWM) weights each individual by their ability (Budescu & Chen, 2015). Specifically, the weights are an individual's ``contribution'' or the degree to which an individual improves the accuracy of the ensemble. If the ensemble accuracy improves when the individual is included, that individual is considered to be a positive contributor. If the accuracy of the ensemble diminishes when the individual is included, that individual is considered to be a negative contributor. In the CWM, only positive contributors are included in the ensemble and those individuals in the ensemble have their individual evaluations weighted by their contribution (i.e., it is a weighted mean).

Because weighted ensemble techniques utilize every individual's accuracy relative to the group, categorical evaluations need to be quantified so that accuracy can be measured. An evaluation's accuracy can be measured using a proper scoring rule like the Brier score (Brier, 1950; Gneiting & Raftery, 2007). Brier scores have two components - the outcome (i.e., whether an event occurred or not) and the probabilistic prediction about whether that event would occur or not. The closer the prediction is to the outcome, the better the accuracy.

CTG evaluations are categorical (e.g., normal, suspicious, pathological). Although FIGO guidelines (Ayres-de-Campos et al., 2015) state that the normal, suspicious, and pathological evaluation categories correspond to no hypoxia, low probability of hypoxia, and high probability of hypoxia respectively, it is not clear what each evaluation category conveys in terms of the probability of a healthy baby given a specific evaluation. There has been research done to derive precise, quantitative estimates of what the probability of hypoxia is for each evaluation category using obstetrician's evaluation data and fetal health outcome (Luhmann & Urs, n.d.). Although we have the probabilistic estimates of what CTG evaluation categories might convey, there has been no investigation into whether using weighted ensemble techniques improves evaluation accuracy amongst obstetricians.

The current study investigates strategies for aggregating obstetricians' CTG evaluations using precise estimates of the probability of hypoxia for each evaluation category. Specifically, our objectives were 1) to determine whether aggregating individual evaluations improves accuracy and 2) to compare the predictive performance of the different aggregation methods. We use the CTU-CHB Intrapartum Cardiotocography Database (Chudaček et al., 2014) which contains partial medical records of 552 mothers and babies, including the CTG recordings, fetal outcomes, and evaluations provided by obstetricians. This was the same database used to derive the quantitative estimates of the evaluation categories in prior work (Luhmann & Urs, n.d.). The aggregation schemes we consider include the CWM described above, its unweighted counterpart, and a simple unweighted average of all obstetricians. These models will be interpreted in comparison to the baseline (performance expected at chance level).

Method

Dataset

The open access intrapartum CTG database (Chudaček et al., 2014) on PhysioNet (Goldberger et al., 2000) was used for the analyses and results reported in the current paper. This database contains 552 intrapartum recordings collected between April 2010 and August 2012 in the Czech Republic and each recording has the relevant biochemical markers (e.g., pH, base excess, base deficit in extracellular fluid) and clinical parameters (e.g., age of the women, sex of the fetus, fetal diseases, gravidity).

Clinical Evaluations

The CTG database includes judgments from nine obstetricians and these evaluations were the basis for the analyses conducted. The obstetricians were from six clinics in the Czech Republic and had delivery experience ranging from 10 to 33 years (with a median value of 15 years). The obstetricians were shown the recordings (CTG tracings) during the various stages of labor.

It is important to note that not all the nine obstetricians evaluated each baby at each step of labor. There were numerous instances in which a given expert evaluated the same baby/step more than once. In such instances, we used the expert's modal judgment. If an obstetrician's judgment did not have a unique mode (e.g., two evaluations of normal and pathological), that obstetrician's judgment for that baby/step were omitted from further analyses. The number of babies evaluated by each expert ranged from 256 to 543 and the number of evaluations per baby varied in each step, ranging from 2 to 9 (step 1), 3 to 9 (step 2), 1 to 9 (steps 3 and 4).

The probability of hypoxia for each evaluation category was calculated via the method used in prior research

(Luhmann & Urs, n.d.). For each obstetrician, a set of three probabilities was obtained with each individual probability corresponding to one of the three evaluation categories - normal, suspicious, and pathological. These three probabilities were estimated independently while enforcing the intuitive ordering. That is, the probability of hypoxia associated with a ``normal'' evaluation was required to be less than or equal to the probability of hypoxia for a ``suspicious'' evaluation, and the probability of hypoxia associated with a ``pathological'' classification was required greater than the probability associated with a ``suspicious'' evaluation. For each obstetrician, the set of probabilities that maximized their accuracy was chosen. In other words, the probabilities of hypoxia were estimated under the assumption that each obstetrician was accurate as possible given their evaluations.

Fetal Outcomes

For all analyses conducted in this study, babies with umbilical pH values less than or equal to 7.05 were considered to be acidotic (i.e., an adverse outcome). Babies with pH values greater than 7.05 were taken as healthy. The threshold of 7.05 has been used in prior research (Costa et al., 2009; Dash et al., 2014), including work by the group responsible for this particular dataset (Abry et al., 2018; Georgoulas et al., 2004; Georgoulas et al., 2017). With this threshold, approximately 93% of the cases in the database were coded as healthy, however, which babies were included in analyses of each step slightly varied (step 1: 92.57% healthy, step 2: 92.64%, step 3: 93.61%, step 4: 95.06%).

Scoring

The Brier score (Brier, 1950) is a proper scoring rule (Gneiting & Raftery, 2007) that can be used to evaluate the performance of probabilistic predictions. When predictions are made on a single dimension (e.g., probability of a healthy fetus), the Brier score reduces to the mean squared difference between the probabilistic prediction and the dichotomous outcome. Thus, Brier scores can range from zero (indicating perfectly correct predictions) to one (indicating perfectly incorrect predictions). In all analyses in this study, the dichotomous outcome of interest was whether a baby was acidotic (i.e., umbilical pH value <=7.05, numerically coded as 0) or not (i.e., umbilical pH value >7.05 and numerically coded as 1).

Aggregation Models

Unweighted Model The unweighted model (UWM) represents the intuitive notion of ensemble forecasts. The ensemble always included all experts and the ensemble evaluation was the simple arithmetic mean of the probabilities associated with the individual obstetrician's evaluations.

Contribution Weighted Models Each expert obstetrician's contribution was calculated separately for each step, by first calculating the Brier score of an ensemble including all the experts that provided an evaluation for that baby at that step of labor. Next, we calculated the Brier score of the same ensemble, but with the target expert omitted. The target obstetrician's contribution was the difference between these two Brier scores.

We investigated two contribution-weighted models. The selection variant of the contribution-weighted model (CWM-S) used contributions to determine which experts were selected into the ensemble. Experts with negative contributions were omitted from the ensemble and evaluations of all positive contributors were combined using a simple arithmetic mean. The continuous variant (CWM-C) also omitted experts with negative contributions, but additionally combined the evaluations of positive contributors via a weighted mean, using their contributions as weights.

Because we are interested in predictive accuracy, we use a standard k-fold cross validation procedure (Hastie et al., 2009), randomly dividing the set of CTG records into five subsets or ``folds''. We then use the records from four of these folds to re-estimate the probability of hypoxia associated with each obstetrician's three evaluations and each obstetrician's contribution and use these to calculate Brier scores using the records in the fifth fold. We then repeated this process five times, using a different fold for scoring each time.

Baseline Model The baseline model (BM) represents the performance expected at chance: the performance expected if one were to give a fixed forecast equal to the empirical base rate of the target outcome. In the current setting, the proportion of healthy babies (i.e., those without an adverse outcome) is approximately 93%, although in each step it varied (step 1: 92.57% healthy, step 2: 92.64%, step 3: 93.61%, step 4: 95.06%). In this way, the baseline does not distinguish among babies and the predictions are thus intentionally ``unskillful'' (Jolliffe & Stephenson, 2008).

Analytic Approach

Each Brier was assumed to be a linear function with a subject-level intercept, two nominal factors (step and model), and noise (error). The intercept was modeled as a random effect. Each record-specific (i.e., fetus-specific) intercept was assumed to be drawn from an overarching normal distribution, which acted as a weakly informative hyper parameter. We used a full rank coding of the categorical variable indicating model (sometimes referred to as index coding), a strategy unavailable in the frequentist framework due to the need for strict identification strategies (rendered unnecessary by the priors in a Bayesian approach). This allows us to calculate arbitrary comparisons (cf. needing to select a specific contrast coding in a frequentist setting). We used reduced rank coding (dummy or treatment coding) of the categorical variable indicating step, with the final step (step 4) acting as the reference. Because Brier scores are bound to the interval between 0 and 1, we conducted our analyses using beta regression (Geissinger et al., 2022), an approach that matches the observed range and naturally handles the expected heteroscedasticity.

All analyses were performed using Bambi (Capretto et al., 2020), a high-level interface for constructing linear models that is built on top of PyMC (Abril-Pla et al., 2023). Bambi automatically generates weakly informative priors for all model parameters, by loosely scaling the priors to the observed data. Posteriors were estimated using the NUTS algorithm (Hoffman et al., 2014). We used four chains, each of which was initialized with 1000 tuning draws (which were then discarded), followed by 1000 draws, for a total of 4000 draws. We observed no divergences, the $\hat{R}$ values were <1.01, and the effective sample sizes were all >650.

Results

Table 1 presents descriptive statistics regarding the Brier Scores for models and Table 2 presents the results of our regression analyses (including means and 94% highest density intervals, HDIs). These results revealed that the Brier scores achieved by the baseline model was inferior when compared to all other models ($p > 0.999$ in all cases). To further investigate, we next compared each pair of models. UWM was more accurate than CWM-C ($p = 0.808$) and UWM was more accurate than CWM-S ($p = 0.784$). Two contribution models were roughly equivalent in their performance (probability that CWM-C was better than CWM-S was 0.546).

We were also interested in how performance varied across steps. Evaluations were more accurate in Step 4 than in Steps 1 and Step 2 ($p = 0.998$ and $p = 1.0$, respectively). Evaluations were also more accurate in Step 3 than Step 2 ($p = 0.956$). Evaluations were also more accurate in Step 2 than Step 1 ($p = 0.703$).

Table 1: Median Brier Scores for models in each step of the labor and delivery process.

Step	CWM-C	CWM-S	UWM	BM
Step 1	0.007	0.007	0.006	0.070
Step 2	0.008	0.008	0.007	0.068
Step 3	0.006	0.005	0.004	0.056
Step 4	0.006	0.004	0.004	0.060

Table 2: Bayesian model results.

	M	94% HDI
Baseline	-1.751	[-1.806, -1.686]
UWM	-2.481	[-2.543, -2.419]
CWM-C	-2.453	[-2.512, -2.388]
CWM-S	-2.457	[-2.515, -2.395]
Step 1	0.093	[0.034, 0.152]
Step 2	0.108	[0.049, 0.168]
Step 3	-0.061	[-0.129, 0.001]

Discussion

Though the use of continuous CTG may provide benefits in the case of certain high-risk pregnancies (Pattison & McCowan, 2004), it has failed to substantially improve fetal outcomes more generally and it has also been shown to substantially increase rates of cesarean deliveries (Alfirevic et al., 2017). In many cases, these cesarean deliveries are intended as an intervention on babies believed to be hypoxic. However, if obstetricians have inflated beliefs about the probability of hypoxia (Luhmann & Urs, n.d.), many of these surgical interventions may be unnecessary (Alfirevic et al., 2017; Nelson et al., 1996), creating risks such as infection, complications for future pregnancies, etc. Despite national and international standards for CTG interpretation (Ayres-de-Campos et al., 2015), the reliability of clinical evaluations is still poor. To overcome this lack of reliability, the current study sought to leverage the so-called ``wisdom of crowds'' (Galton, 1907; Hertwig, 2012; Surowiecki, 2005), reducing noise in individual judgments and revealing the predictive skill common across the individual judgments (Satopää et al., 2014).

In order to aggregate obstetricians' judgments, the categorical evaluations (normal, suspicious, pathological) needed to be transformed into precise, quantitative estimates of the probability of hypoxia for a baby in each evaluation category for each obstetrician. The FIGO guidelines indicate no, low, and high probability of hypoxia to the normal, suspicious, and pathological evaluation categories respectively, but it is unclear how each obstetrician is using the guidelines and what their implied probabilities for the evaluation categories are. Therefore, we utilized the method in prior research (Luhmann & Urs, n.d.) to get optimized estimates of the probability of hypoxia. We generated sets of three probabilities of hypoxia for each obstetrician, where each probability represented the probability of hypoxia for a baby given a specific evaluation (i.e., normal, suspicious, pathological). For each obstetrician, the set of three probabilities that maximized accuracy was selected. The advantage of this approach is that the probability of hypoxia one obstetrician might associate with a normal evaluation may be different from other obstetricians' probability association with a normal evaluation category. Therefore, the probabilities associated are obstetrician-specific, which may be the case in real-life where individual differences in training, risk attitudes, and experience may shape how an obstetrician evaluates babies' risk of hypoxia.

We used three techniques to aggregate the obstetricians' evaluations (i.e., the optimized estimates of the probability): an unweighted model which generates the mean of all the obstetricians, a contribution model where only the unweighted mean of the positive contributors' judgments is obtained, and a contribution model where the weighted mean of the positive contributors' is obtained. We compared all three aggregation models to a baseline model where the output was always the base rate of hypoxia in the dataset for each step of the labor and delivery process.

We find that aggregating the judgments of obstetricians enhances accuracy, aligning with the ``wisdom of crowds'' phenomenon (Budescu & Chen, 2015; Galton, 1907; Surowiecki, 2005). However, simple averaging outperformed

the more sophisticated methods. There was no additional benefit of aggregating the judgments of obstetricians who were positive contributors to their ensemble via simple or weighted averaging. Regardless, across all the aggregation models, accuracy improved over the course of labor, suggesting that obstetricians' evaluations may be able to better contextualize the CTG tracings and provide more accurate evaluations as more information about the baby is revealed.

Aggregation seems to improve obstetricians' evaluations and may serve as a solution to the reliability and accuracy issues that persist in fetal heart rate evaluation. It must be noted that the aggregation methods relied on optimized estimates of the probability of hypoxia in each evaluation category. The aggregation of evaluation improves accuracy under the assumption that we are using the best interpretations of what each obstetrician means in terms of the probability of hypoxia for each evaluation category. As a consequence, use of these techniques in the real world would require collecting the evaluations of obstetricians over a period of time as well as collecting information about the evaluated babies' umbilical pH. Our approach of aggregating obstetricians' evaluations in order to improve accuracy is viable as long as there is a history of obstetricians' evaluations and fetal health outcomes but may not be useful for live-decision optimization. Our use of optimized estimates of the probability of hypoxia for the evaluation classifications requires a prior base of evaluations and obstetricians' contributions to the ensemble can only be computed using historical data.

The finding that aggregating judgments of obstetricians improves accuracy align with findings from other researchers on the wisdom of crowds (Galton, 1907; Surowiecki, 2005). But the lack of improvement in accuracy in the CWM-S and CWM-C models (models which accounted for the obstetricians' contributions in an ensemble) was not in line with those of researchers demonstrated who demonstrated that contribution-based models outperformed simple averaging (Budescu & Chen, 2015). One reason as to why the UWM may have outperformed CWM-S and CWM-C is that because no one was removed from the ensemble in UWM, a larger number of judgments were aggregated. When a larger set of judgments is aggregated, the signal-to-noise ratio is optimized and evidence shows that aggregating judgments of larger groups tends to produce greater accuracy (Bassamboo et al., 2015; Brown & Yang, 2019; Huber & Delbecq, 1972; Walter et al., 2022). In addition, we lack full information about obstetricians' background which limit explanations based on skill level and training received.

Aggregating the judgments of large groups of people in order to make high-quality decisions has been in practice in numerous domains. In business, companies run internal prediction markets which pool predictions in order to determine whether a project was worth undertaking (Borison & Hamm, 2010; Cowgill & Zitzewitz, 2015). The level of bias tends to get lower as the number of people who participate in these markets increase. In the land development sector, researchers have proposed combining individuals' judgment in public participation in information systems to improve public participation in planning (Brown, 2015). Based on the findings in this study, one practical application in the domain of obstetrics may be to obtain and aggregate more judgments of obstetricians or even other healthcare providers (for example nurses and midwives) in order to reduce the bias during decision-making and improve accuracy in fetal outcome evaluations. Although the exact techniques cannot be used in live-decision making in the delivery room, it provides evidence that getting as many judgments as possible may help with accuracy. In the delivery room, the weights placed on obstetricians' judgments increase by seniority. In our study, it was not possible to obtain this information to include in our models but future research can investigate alternative weighting strategies that may reflect real-life scenarios.

References

Abril-Pla, O., Andreani, V., Carroll, C., Dong, L., Fonnesbeck, C., Kochurov, M., Kumar, R., Lao, J., Luhmann, C., Martin, O., & others (2023). PyMC: a modern, and comprehensive probabilistic programming framework in Python. *PeerJ Computer Science, 9*, e1516.

Abry, P., Spilka, J., Leonarduzzi, R., Chudáček, V., Pustelnik, N., & Doret, M. (2018). Sparse learning for Intrapartum fetal heart rate analysis. *Biomedical Physics & Engineering Express, 4(3), 034002.*

Alfirevic, Z., Gyte, G., Cuthbert, A., & Devane, D. (2017). Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour. *Cochrane database of systematic reviews (2).*

American College of Obstetricians, Gynecologists, & others (2010). Practice bulletin no. 116: Management of intrapartum fetal heart rate tracings. *Obstetrics and gynecology, 116*(5), 1232–1240.

Ayres-de-Campos, D., Spong, C., Chandraharan, E., & others (2015). FIGO consensus guidelines on intrapartum fetal monitoring: Cardiotocography. *International Journal of Gynecology & Obstetrics, 131*(1), 13–24.

Barnett, M., Boddupalli, D., Nundy, S., & Bates, D. (2019). Comparative accuracy of diagnosis by collective intelligence of multiple physicians vs individual physicians. *JAMA network open, 2*(3), e190096–e190096.

Bassamboo, A., Cui, R., & Moreno, A. (2015). The wisdom of crowds in operations: Forecasting using prediction markets. SSRN Electron. J.

Bernardes, J., Costa-Pereira, A., Ayres-de-Campos, D., Van Geijn, H., & Pereira-Leite, L. (1997). Evaluation of interobserver agreement of cardiotocograms. *International Journal of Gynecology & Obstetrics, 57(1), 33–37.*

Bernstein, J., Long, J., Veillette, C., & Ahn, J. (2011). Crowd intelligence for the classification of fractures and beyond. *PLoS One, 6*(11), e27620.

Blackwell, S., Grobman, W., Antoniewicz, L., Hutchinson, M., & Bannerman, C. (2011). Interobserver and

intraobserver reliability of the NICHD 3-tier fetal heart rate interpretation system. *American journal of obstetrics and gynecology, 205*(4), 378–e1.

Blix, E., Sviggum, O., Koss, K., & Oian, P. (2003). Inter-observer variation in assessment of 845 labour admission tests: comparison between midwives and obstetricians in the clinical setting and two experts.*BJOG: an international journal of obstetrics and gynaecology, 110*(1), 1–5.

Blutinger, E., Shahid, S., Jarou, Z., Schneider, S., Kang, C., & Rosenberg, M. (2021). Translating COVID-19 knowledge to practice: Enhancing emergency medicine using the "wisdom of crowds". *Journal of the American College of Emergency Physicians Open, 2*(1), e12356.

Borison, A., & Hamm, G. (2010). Prediction Markets: A New Tool for Strategic Decision Making. *California Management Review, 52(4), 125–141.*

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review, 78*(1), 1–3.

Brown, G. (2015). Engaging the wisdom of crowds and public judgement for land use planning using public participation geographic information systems. *Australian Planner, 52(3), 199–209.*

Brown, A., & Yang, F. (2019). The wisdom of large and small crowds: Evidence from repeated natural experiments in sports betting. *Int. J. Forecast., 35(1), 288–296.*

Budescu, D., & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management science, 61*(2), 267–280.

Capretto, T., Piho, C., Kumar, R., Westfall, J., Yarkoni, T., & Martin, O. (2020). Bambi: A simple interface for fitting Bayesian linear models in Python. *arXiv preprint arXiv:2012.10754*.

Chudaček, V., Spilka, J., Burša, M., Janku, P., Hruban, L., Huptych, M., & Lhotská, L. (2014). Open access intrapartum CTG database. *BMC pregnancy and childbirth, 14*, 1–12.

Costa, A., Ayres-de-Campos, D., Costa, F., Santos, C., & Bernardes, J. (2009). Prediction of neonatal acidemia by computer analysis of fetal heart rate and ST event signals. *American journal of obstetrics and gynecology, 201*(5), 464–e1.

Cowgill, B., & Zitzewitz, E. (2015). Corporate prediction markets: Evidence from Google, ford, and firm X. *Rev. Econ. Stud., 82(4), 1309–1341.*

Dash, S., Quirk, J., & Djuric, P. (2014). Fetal heart rate classification using generative models. *IEEE Transactions on Biomedical Engineering, 61*(11), 2796–2805.

Declercq, E., Sakala, C., Corry, M., Applebaum, S., & Herrlich, A. (2014). Major survey findings of Listening to MothersSM III: Pregnancy and Birth. *The Journal of perinatal education, 23*(1), 9–16.

Devane, D., & Lalor, J. (2005). Midwives' visual interpretation of intrapartum cardiotocographs: intra-and inter-observer agreement. *Journal of advanced nursing, 52*(2), 133–141.

Dhami, M., Mandel, D., Mellers, B., & Tetlock, P. (2015). Improving intelligence analysis with decision science. *Perspectives on Psychological Science, 10*(6), 753–757.

Donker, D., Geijn, H., & Hasman, A. (1993). Interobserver variation in the assessment of fetal heart rate recordings. *European Journal of Obstetrics & Gynecology and Reproductive Biology, 52*(1), 21–28.

Galton, F. (1907). Vox Populi. Nature, 75(1949), 450–451.

Georgoulas, G., Stylios, C., Nokas, G., & Groumpos, P. (2004). Classification of fetal heart rate during labour using hidden Markov models. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)* (pp. 2471–2475).

Geissinger, E., Khoo, C., Richmond, I., Faulkner, S., & Schneider, D. (2022). A case for beta regression in the natural sciences. *Ecosphere, 13*(2), e3940.

Georgoulas, G., Karvelis, P., Spilka, J., Chudáček, V., Stylios, C., & Lhotská, L. (2017). Investigating pH based evaluation of fetal heart rate (FHR) recordings. *Health and technology, 7, 241–254.*

Gneiting, T., & Raftery, A. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association, 102*(477), 359–378.

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., Mietus, J., Moody, G., Peng, C.K., & Stanley, H. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation, 101(*23), e215–e220.

Graham, E., Petersen, S., Christo, D., & Fox, H. (2006). Intrapartum electronic fetal heart rate monitoring and the prevention of perinatal brain injury. *Obstetrics & Gynecology, 108(3 Part 1), 656–666.*

Grant, A., Joy, M.T., O'Brien, N., Hennessy, E., & Macdonald, D. (1989). Cerebral palsy among children born during the Dublin randomised trial of intrapartum monitoring. *The Lancet, 334*(8674), 1233–1236.

Green, S., Martinez-Rumayor, A., Gregory, S., Baggish, A., O'Donoghue, M., Green, J., Lewandrowski, K., & Januzzi, J. (2008). Clinical uncertainty, diagnostic accuracy, and outcomes in emergency department patients presenting with dyspnea. *Archives of Internal Medicine, 168*(7), 741–748.

Hastie, T., Tibshirani, R., Friedman, J., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. (Vol. 2) Springer.

Hermann, B., Goudard, G., Courcoux, K., Valente, M., Labat, S., Despois, L., Bourmaleau, J., Richard-Gilis, L., Faugeras, F., Demeret, S., & others (2019). Wisdom of the caregivers: pooling individual subjective reports to diagnose states of consciousness in brain-injured patients, a monocentric prospective study. *BMJ open, 9*(2).

Hertwig, R. (2012). Tapping into the wisdom of the crowd—with confidence. Science, 336(6079), 303–304.

Hoffman, M., Gelman, A., & others (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research, 15(1), 1593–1623.*

Hruban, L., Spilka, J., Chudáček, V., Janku, P., Huptych, M., Burvsa, M., Hudec, A., Kacerovsky, M., Koucky, M., Procházka, M., & others (2015). Agreement on intrapartum cardiotocogram recordings between expert obstetricians. *Journal of evaluation in clinical practice, 21*(4), 694–702.

Huber, G., & Delbecq, A. (1972). Guidelines for Combining the Judgments of Individual Members in Decision Conferences. *Academy of Management Journal, 15(2), 161–174.*

Jolliffe, I., & Stephenson, D. (2008). Proper scores for probability forecasts can never be equitable. *Monthly weather review, 136*(4), 1505–1510.

Kämmer, J., Hautz, W., Herzog, S., Kunina-Habenicht, O., & Kurvers, R. (2017). The potential of collective intelligence in emergency medicine: pooling medical students' independent decisions improves diagnostic performance. *Medical decision making, 37(6), 715–724.*

Kattan, M., O'Rourke, C., Yu, C., & Chagin, K. (2016). The wisdom of crowds of doctors: Their average predictions outperform their individual ones. *Medical Decision Making, 36*(4), 536–540.

Krockow, E., Kurvers, R., Herzog, S., Kämmer, J., Hamilton, R., Thilly, N., Macheda, G., & Pulcini, C. (2020). Harnessing the wisdom of crowds can improve guideline compliance of antibiotic prescribers and support antimicrobial stewardship. *Scientific reports, 10*(1), 18782.

Kurvers, R., Herzog, S., Hertwig, R., Krause, J., Carney, P., Bogart, A., Argenziano, G., Zalaudek, I., & Wolf, M. (2016). Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences, 113*(31), 8777–8782.

Luhmann, C., & Urs, M. (Manuscript submitted for publication). The Quantitative Meaning of CTG Evaluations and their Predictive Accuracy.

Molteni, F., Buizza, R., Palmer, T., & Petroliagis, T. (1996). The ECMWF ensemble prediction system: Methodology and validation. *Quarterly journal of the royal meteorological society, 122(*529), 73–119.

Nelson, K., Dambrosia, J., Ting, T., & Grether, J. (1996). Uncertain value of electronic fetal monitoring in predicting cerebral palsy. *New England Journal of Medicine, 334(10), 613–619.*

Nielsen, P., Stigsby, B., Nickelsen, C., & Nim, J. (1987). Intra-and inter-observer variability in the assessment of intrapartum cardiotocograms. *Acta obstetricia et gynecologica Scandinavica, 66(5), 421–424.*

O'Brien-Abel, N. (2020). Clinical implications of fetal heart rate interpretation based on underlying physiology. MCN: *The American Journal of Maternal/Child Nursing, 45*(2), 82–91.

Palomäki, O., Luukkaala, T., Luoto, R., & Tuimala, R. (2006). Intrapartum cardiotocography–the dilemma of interpretational variation. *Journal of Perinatal Medicine*.

Pattison, N., & McCowan, L. (2004). Cardiotocography for antepartum fetal assessment (Cochrane Review). *The Cochrane Library, 2.*

Poses, R., Bekes, C., Winkler, R., Scott, W., & Copare, F. (1990). Are Two (Inexperienced) Heads Better Than One (Experienced) Head?: Averaging House Officers' Prognostic Judgments for Critically Ill Patients. *Archives of internal medicine, 150*(9), 1874–1878.

Satopäa, V., Baron, J., Foster, D., Mellers, B., Tetlock, P., & Ungar, L. (2014). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting, 30*(2), 344–356.

Sholapurkar, S. (2015). Categorization of fetal heart rate decelerations in American and European practice: importance and imperative of avoiding framing and confirmation biases. Journal of clinical medicine research, 7(9), 672.

Shy, K., Luthy, D., Bennett, F., Whitfield, M., Larson, E., Van Belle, G., Hughes, J., Wilson, J., & Stenchever, M. (1990). Effects of electronic fetal-heart-rate monitoring, as compared with periodic auscultation, on the neurologic development of premature infants. *New England Journal of Medicine, 322(9), 588–593.*

Spilka, J., Chudáček, V., Janku, P., Hruban, L., Burvsa, M., Huptych, M., Zach, L., & Lhotská, L. (2014). Analysis of obstetricians' decision making on CTG recordings. *Journal of biomedical informatics, 51, 72–79.*

Surowiecki, J. (2005). The wisdom of crowds. Anchor Books.

Tetlock, P., & Gardner, D. (2016). Superforecasting. Broadway Books.

Todros, T., Preve, C., Plazzotta, C., Biolcati, M., & Lombardo, P. (1996). Fetal heart rate tracings: observers versus computer assessment. *European Journal of Obstetrics & Gynecology and Reproductive Biology, 68*, 83–86.

Toth, Z., & Kalnay, E. (1993). Ensemble forecasting at NMC: The generation of perturbations. *Bulletin of the American meteorological society, 74*(12), 2317–2330.

Trimbos, J., & Keirse, M. (1978). Observer variability in assessment of antepartum cardiotocograms. *BJOG: An International Journal of Obstetrics & Gynaecology, 85(12), 900–906.*

Walter, V., Kölle, M., & Collmar, D. (2022). Measuring the Wisdom of the Crowd: How Many is Enough?. *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science, 90(3), 269–291.*

Windrix, C. (2017). Fetal Heart Rate Monitoring. Data Interpretation in Anesthesia: A Clinical Guide, 73–75.